

Genome analysis

Simulation based estimation of branching models for LTR retrotransposons

Serge Moulin^{1*}, Nicolas Seux², Stéphane Chrétien³, Christophe Guyeux¹, and Emmanuelle Lerat⁴,

¹Département d'Informatique des Systèmes Complexes, UMR 6174 CNRS, FEMTO-ST Institute, 15 Bis Avenue des Montboucons, 25030 Besançon, France

²Laboratoire de Mathématiques, Université de Franche-Comté, UMR 6623 CNRS, 16 route de Gray, 25030 Besançon, France

³National Physical Laboratory, Hampton Road, Teddington, United Kingdom and

⁴Laboratoire Biometrie et Biologie Evolutive, Université Claude Bernard - Lyon 1, UMR 5558 CNRS, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne, France

Associate Editor: XXXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: LTR retrotransposons are mobile elements that are able, like retroviruses, to copy and move inside eukaryotic genomes. In the present work, we propose a branching model for studying the propagation of LTR retrotransposons in these genomes. This model allows to take into account both positions and degradations of LTR retrotransposons copies. In our model, the duplication rate is also allowed to vary with the degradation level.

Results: Various functions have been implemented in order to simulate their spread and visualization tools are proposed. Based on these simulation tools, we show that an accurate estimation of the parameters of this propagation model can be performed. We applied this method to the study of the spread of the transposable elements ROO, Gypsy, and DM412 on a chromosome of *Drosophila melanogaster*.

Availability: Our proposal has been implemented using Python software. Source code is freely available on the web at <https://github.com/SergeMOULIN/retrotransposons-spread>.

Contact: serge.moulin@univ-fcomte.fr

1 Introduction

A transposable element (TE) is a DNA sequence able to move from one location to another inside a genome. These sequences, discovered during the 50's by Barbara McClintock (McClintock, 1987) exist in almost all living organisms and are the source of a huge number of mutations. They are considered as a major cause of genetic disease in human (Belancio *et al.*, 2008) or in *Drosophila* where they are responsible for more than 80% of the spontaneous mutations (Green, 1988). DNA sequences derived from these TEs can represent a large part of a genome. For example they represent about 45% of the human genome (Lander *et al.*, 2001) and over 70% of the corn genome (Sanmiguel and Bennetzen, 1998). Hopefully, most of these sequences correspond to fragments or "dead" elements that have lost their ability to move in the genome due to several lethal mutations or are controlled especially via epigenetic mechanisms.

TEs have two possible ways to move in a genome, according to their type (Wicker *et al.*, 2007). The first class of mobile elements are cut from their original place to move to another one, and are called "DNA transposons" or "Class II transposable elements". The other class of mobile elements, called "retrotransposons" or "Class I transposable elements", use an RNA intermediate to duplicate themselves, the new copy being inserted into another location of the genome. Two orders are identified among the retrotransposons according to the presence or absence of Long Terminal Repeat (LTR) sequences at their extremities. The LTR retrotransposons are similar in structure to retroviruses such as HIV. In both classes, TEs can be classified as either "autonomous", if they code the enzymes that will allow them to move, or "non autonomous" if they use the enzymes produced by other elements. In an assembled genome, the various sequences corresponding to TE insertions can be found using different bioinformatic approaches (see (Lerat *et al.*, 2011) for a review), which allow to determine the exact number and positions of each TE insertion. In this article, we focused on the important problem of inferring the history

of the spreading of LTR retrotransposons. For this purpose, we modeled the evolution using a branching process where each element (*i.e.*, a copy of some TE) can randomly evolve via duplication or mutation.

Instances of branching processes have been proposed in the literature as models for the propagation of TEs. To the best of our knowledge, the first model of that kind has been designed by Kaplan *et al.* in 1985 (Kaplan *et al.*, 1985). In this paper, the authors proposed a model where TEs can be either of wild type (*i.e.*, non mutated) or of mutant type. At each generation, wild copies can mutate or disappear, whereas mutant ones can only disappear, and the number of new copies created by transposition at each generation is assumed to follow a Poisson distribution. This number is supposed to decrease with the number of surviving copies, in order to account for the reducing number of unoccupied sites where transposition could take place. This number of new copies is also assumed to be a decreasing function of the proportion of mutants. Finally, the proportion of mutant type among newly created copies increases with the proportion of mutant type in existing copies. The temporal behavior of the number of copies of each type has been studied as well as the extinction time of the TE families.

In 1988, Michael E. Moody (Moody, 1988) has used a branching model to describe the propagation of TEs in a haploid population. The variable studied was the number of individuals possessing i copies of a given TE (denoted by $Z_t^{(i)}$ for the t^{th} generation). This number of copies was itself limited by the number m of viable sites where TE could settle down. This model incorporated a copy-dependent selection. That is to say, the number of identical offspring produced by an individual is supposed to follow a law $Y^{(i)}$ dependent on i . In addition, some probabilities were defined, p_i , q_i , and $1 - p_i - q_i$ for an individual to: gain a copy of the TE, lose one element, or remain stable ($q_m = 0$). Another probability was defined β for an individual devoid of the element to acquire a single copy. Finally, the asymptotic stability of $Z_t = (Z_t^{(0)}, \dots, Z_t^{(max)})$ was studied based on these factors. This study shows that the TE load can maintain itself at intermediate level even in case of a neutral selection. In particular, models with copy-dependent transposition rate can provide a stationary distribution when the deletion rate is between the extremes of transposition rate.

Sawyer *et al.* (Sawyer *et al.*, 1987) produced a model very close to Moody’s in order to study the distribution and abundance of insertion sequences, which are DNA transposons, among natural isolates of *Escherichia coli*. The main difference is that deletion is not taken into account. The reason for this choice is that the data available seemed to indicate that deletion occurs at a substantially smaller rate than transposition (Egner and Berg, 1981; Foster *et al.*, 1981).

More recently, interesting models have been proposed that take into account the location of TEs. For instance, Drakos and Wahl (Drakos and Wahl, 2015) suggested a model of mobile promoter evolution where the probabilities for promoters to duplicate inside or outside their region were potentially different.

In the present work, the objective is to combine a location dependent model with the fact that LTR retrotransposons can face degradation, which may decrease their duplication rate. This model assumes that each copy can be either duplicated or mutated at any time. The time before duplication is supposed to follow an exponential distribution, as well as the time before mutation, the impact of mutation, and the distance covered by a duplicated copy before insertion. This model also takes into account the position of the root (*i.e.* the first copy), and the impact of mutations faced by a TE on its duplication speed. Our main goal is to estimate the parameters for this model. For this purpose, we propose a simulation based procedure. This method requires to define a distance between the results of the simulations and the observed genome, which is based on the Hungarian method (Kuhn, 1955; Munkres, 1957). This method has been applied to the spreading of the LTR retrotransposons ROO, DM412, and Gypsy on the chromosome 3L of *Drosophila melanogaster*. The parameters associated to each TE

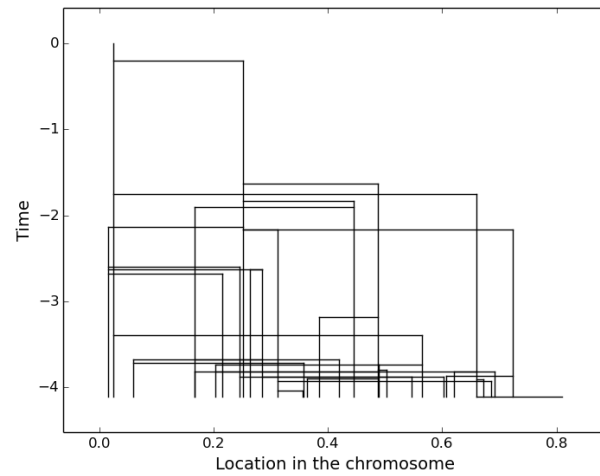


Fig. 1. ROO spread

are computed and a branching tree is proposed in each case. Our results show that, according to our model and method, the roots of ROO, DM412, and Gypsy on the chromosome 3L could correspond respectively to the annotated copies FBti0060418, FBti0020009, and FBti0020033.

2 System and methods

2.1 The branching model

2.1.1 The branching tree

An example of branching tree is shown in Figure 1.

This branching tree represents the spread of the LTR retrotransposons “ROO” between times 0 and 4.105. At time 0 there is only one copy, called the “root” in this article. At time 0.205, the root duplicates itself to give birth to its first “child”. Finally, the genome is observed at time $T_{obs} = 4.105$ with 32 copies of “ROO” inserted in it. The state of the tree at time $T_{obs} = 4.105$ is named “final state” of this tree. The working principle of our estimation method is to simulate trees in order to determine in which conditions final states of simulated trees match well with the observed genome. The branching tree represents only duplications, but copies are also subject to mutations as explained in Section 2.1.2. To compare the final state of a simulated tree with the observed genome, the copy locations in the genome have been considered as well as their mutations, as detailed in Section 2.2.1.

2.1.2 The general model

The branching model is constructed as follows.

1. The spread starts with only one copy, called the root, at time zero in a location X_0 to determine.
2. Each copy can be either duplicated or mutated at any time.
3. The number of copies increases due to duplications. When a new copy is created, it receives an index equal to the number of existing copies at the time of its birth, including itself. In the remainder of this article, let T_i be the birth date of the i^{th} copy and let $\tau_{i,k}$ be the time when the i^{th} copy faces its k^{th} mutation.
4. The time intervals $\tau_{i,k+1} - \tau_{i,k}$ between two mutations is supposed to be independent and identically distributed (i.i.d.) with exponential distribution $\mathcal{E}(\frac{1}{\mu})$, where μ (*i.e.*, average time between two mutations) must be determined.

5. The proportion of nucleotides affected by a mutation is supposed to follow a distribution $\min(1, \mathcal{E}(\frac{1}{\beta}))$, where β must be determined.
6. Each copy is also associated to its Needleman-Wunsch (Needleman and Wunsch, 1970) distance to the original state of the root. This distance, also named “state of deterioration” in the remainder of this article, is denoted by $D_i(t)$ for the i^{th} copy at time t . This distance increases as a function of time, due to mutation effects. Finally, $D_i(T_{obs})$ is the state of deterioration at the end of the spread.
7. At time t , for the i^{th} copy, conditionally on $D_i(t)$, we assume that the time before the next duplication follows a distribution $\mathcal{E}(\frac{1}{1+p \times D_i(t)})$, where $p > 0$ is a parameter to determine. In other words, the time before the next duplication is longer when the copy is far from the original state of the root in terms of Needleman-Wunsch distance.
8. Moreover, each copy is also associated to its position in the genome. This position is denoted by X_i for the i^{th} copy. This position is constant with time. We assume that each child j of a copy i satisfies $X_j = X_i + \chi_{i,j}$, where $\chi_{i,j}$ follows a distribution $\mathcal{U}\{-1, 1\} \times \mathcal{E}(\frac{1}{L})$, in which \mathcal{U} represents the uniform law (*i.e.*, the probability to choose -1 or 1 is the same) and L is a parameter to determine.

Our goal is thus to estimate the parameters of this model, *i.e.*, X_0, μ, β, p , and L . Note that the duplication speed of the non-mutated root is set to 1 and it does not need to be determined. Indeed, this duplication speed is redundant with μ and p .

2.2 The estimation method

As explained in Section 2.1.1, the working principle of our estimation method is to simulate trees in order to determine in which conditions the final states of simulated trees match well with the observed genome.

Trees are simulated according to the model defined in Section 2.1.2. The stopping criterion of these simulated trees depends on the number of copies in the observed genome. Actually the simulation was constrained to stop randomly between the birth date of the n^{th} and the birth date of the $n + 1^{th}$ copy, where n is the number of copies in the observed genome, see Section 3.1.2 for further details.

In what follows, Section 2.2.1 defines the distance between the final state of a simulated tree and the observed genome, and Section 2.2.2 presents how this distance has been minimized.

2.2.1 Distance between trees

The first step is to define a distance between the final state of a simulated tree and the observed genome, that is to say, a distance between final states of two trees. For this purpose, a distance between two copies has been designed as follows.

$$D(R_i, \tilde{R}_j) = |X_i - \tilde{X}_j| + |D_i(T_{obs}) - \tilde{D}_j(T_{obs})|$$

where R_i is the i^{th} copy of the first tree and \tilde{R}_j the j^{th} copy of the second one. $|X_i - \tilde{X}_j|$ represents here the geographical distance between R_i and \tilde{R}_j while $D_i(T_{obs})$ is for the Needleman-Wunsch distance between R_i and the state of the root at the origin. Then, the distance between final states of two trees has been defined as the best possible adjustment between copies, using Kuhn-Munkres algorithm (also named Hungarian method) (Kuhn, 1955; Munkres, 1957).

To illustrate this distance, let us consider the following example: at its final state, the first tree has 3 copies of a given TE. Their respective locations in the genome are $X_1 = 0.58$, $X_2 = 0.03$, and $X_3 = 0.97$ while their respective Needleman-Wunsch distances from the original state of the root are $D_1(1) = 0.04$, $D_2(1) = 0.13$, and $D_3(1) = 0.14$. The second tree has also 3 copies. Their respective locations in the genome are $\tilde{X}_1 = 0.55$, $\tilde{X}_2 = 0.90$, and $\tilde{X}_3 = 0.96$ while their respective Needleman-Wunsch

		Table 1. The matrix of distance between copies		
		$X_1 = 0.58$	$X_2 = 0.03$	$X_3 = 0.97$
		$D_1(T_{obs}) = 0.04$	$D_2(T_{obs}) = 0.13$	$D_3(T_{obs}) = 0.14$
$\tilde{X}_1 = 0.55$	$\tilde{D}_1(T_{obs}) = 0.07$	0.06	0.58	0.49
$\tilde{X}_2 = 0.90$	$\tilde{D}_2(T_{obs}) = 0.06$	0.34	0.94	0.15
$\tilde{X}_3 = 0.96$	$\tilde{D}_3(T_{obs}) = 0.08$	0.42	0.98	0.07

distances from the original state of the root are $D_1(1) = 0.07$, $D_2(1) = 0.06$, and $D_3(1) = 0.08$.

The matrix of distances between copies of the two trees can thus be constructed as in Table 1.

According to the Kuhn-Munkres algorithm, the best assignment is \tilde{R}_1 with R_2 , \tilde{R}_2 with R_1 , and \tilde{R}_3 with R_3 . The distance between final states of these two trees is thus $0.58 + 0.34 + 0.07 = 0.99$. For the remainder of this article, let D_T be the distance between two trees as defined in this section.

2.2.2 Minimisation of the distance

The objective now is to determine the parameter set (X_0, μ, β, p , and L) that minimizes in average the distance between the simulated trees and the observed genome. For this purpose, a 5-dimensional grid has been constructed, where each point of this grid represents a parameter set.

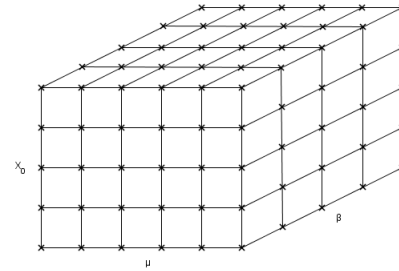


Fig. 2. A parameter set grid

A first score $S_1 = \sum_{i=1}^{N1} D_T(T_i, G)$ has been associated to each of these parameter sets. In this formula, T_i is the i^{th} tree simulated with the parameter set, G represents the observed genome, and $N1$ is a parameter chosen by the user of the optimization method (for instance $N1 = 100$ for the case study in Section 4). The best $N2$ points have then been stored and a new score $S_2 = \sum_{i=1}^{N3} D_T(T_i, G)$ has been associated to each of them. $N2$ and $N3$ are also parameters of the optimization method, with the constraints that $N2$ is lower than the number of points of the grid and that $N3$ is larger than $N1$. Finally the best of these points are selected to construct a smaller grid around it. This iterative process is continued until the precision chosen by the user of the optimization method has been obtained.

Table 2. Example of the output T

[0.5	0.	0.]
[0.19031606	1.83699228	0.]
[0.18321005	11.25706728	0.]
[0.66442132	17.61532334	2.]
[0.48479738	25.45993783	1.]
[0.13876928	28.11662473	1.]

3 Algorithm

Our proposal has been implemented using Python¹. A short application programming interface is detailed thereafter.

3.1 TreeBuild

This main function is used to build branching trees following the model defined in Section 2.1.2. Its halt condition is the targeted number of copies, while its prototype meets the following canvas:

$$(S, T, time) = TreeBuild(X, mu, B, p, L, n),$$

where n is the desired number of copies, while X , μ , β , p , and L are the model parameters as defined in Section 2.1.2. S is a $n \times 1$ vector representing states of deterioration. $time$ is the propagation time, also named T_{obs} in this article. Finally T is a $n \times 3$ matrix containing, for each copy: its position, its birth date, and the row of its mother, like in Table 2.

In this example, the mother of the copy located in 0.18 is the root. The mother of the copy located in 0.66 is the copy located in 0.18. Other details regarding this main function are provided thereafter.

3.1.1 Multiple clocks management

The working principle of TreeBuild can be summarized as follows: it determines the next event (mutation or duplication) and executes it until the stopping criterion is satisfied. To determine the next event means to know its nature (mutation or duplication), its time, and in which of the available copies it happens. Let j be the number of available copies at time t_1 . The easiest way to determine the next event is to simulate $2 \times j$ exponential laws, one for each possible mutation or duplication. The minimum of these $2 \times j$ simulations can thus provide the time, the nature, and the copy related to the next event.

Actually, TreeBuild does not really simulate $2 \times j$ exponential laws, as two properties of this law have been used to shorten computations. Indeed, $\forall (p_1, \dots, p_{2j}) \in \mathbb{R}^{2j}, \forall (Y_1, \dots, Y_{2j}) \sim (\mathcal{E}(p_1), \dots, \mathcal{E}(p_{2j}))$, we have:

1. $\min(Y_1, \dots, Y_{2j}) \sim \mathcal{E}\left(\sum_{i=1}^{2j} p_i\right)$,
2. $\forall i \in 1 \dots 2j, P(Y_i = \min(Y_1, \dots, Y_{2j})) = \frac{p_i}{\sum_{k=1}^{2j} p_k}$.

Hence, due to the first property, the time of the next event can be simulated by a single exponential law. The second property, for its part, allows to determine the nature and the copy affected by the next event using a single uniform law.

3.1.2 Stopping criterion

As stated before, the stopping criterion of TreeBuild is related to n (the number of copies of the observed genome). But when a genome is observed, there is no way to detect that a new duplication has just

occurred. Thus, the program cannot stop exactly at the birth of the n^{th} copy. Actually, TreeBuild must run until the T_{n+1} iteration (the birth date of the $n + 1^{th}$ copy), and then the propagation time T_{obs} can be determined by: $T_{obs} = T_n + U \times (T_{n+1} - T_n)$, where $U \sim U[0, 1]$.

Furthermore, each value taken by S and T between T_n and T_{n+1} is kept in memory. Thereby, the values of T and S returned by TreeBuild are values of T and S at time T_{obs} .

3.1.3 The management of copy locations

Copy positions in the chromosome are in the interval $[0, 1]$. The distance traveled by a TE before insertion is assumed to follow an exponential law, but this latter can send the new copy outside the interval $[0, 1]$. The solution chosen in this case is to launch again the computation of the new copy position. In other terms:

$$\begin{aligned} &\text{While } (X_{child} \notin [0, 1]) : \\ &\quad X_{child} = X_{mother} + U \times Y \end{aligned}$$

where $U \sim U\{-1, 1\}$ and $Y \sim \mathcal{E}(\frac{1}{L})$.

3.1.4 Critical situations

When TreeBuild is launched for each point of the grid of parameters, some critical situations can happen, which may induce a significant slowdown of the program. In particular, when μ is small, and β is large, the probability that an event will be a duplication rather than a mutation becomes very low. Thus, TreeBuild executes an inordinate number of mutations before reaching the desired number of copies. Furthermore, the state of deterioration of a copy that faces a great number of mutations converges towards 0.75. To solve this issue, we have decided that, when the state of deterioration of a copy is in $[0.74, 0.76]$, then this copy cannot mutate anymore.

3.2 Estimation method

In the available package, the estimation of the branching model parameters is realized by the *Optim* function. Its prototype is as follows:

$$Best = optim(Grid, Case, n1, N1, N2, N3).$$

Here, *Grid* is a 5×4 matrix of settings defined exactly as in Section 3. *Case*, for its part, is a $2 \times n$ matrix containing locations and state of deterioration for each copy of the observed genome. $N1$, $N2$, and $N3$ are settings defined in Section 2.2.2 and $n1$ indicates how the grid is shrunk at each step after obtaining the best point (cf. the following section). The output *Best* is the parameter set (X_0, μ, β, p, L) returned by the *Optim* function.

3.2.1 Interval reduction

As explained in Section 2.2.2, the estimation method works with a grid where each point represents a parameter set. When the best point of the grid is found, a new grid is constructed around this point. Note that the new grid is not necessarily included in the previous one, in order to provide a larger degree of freedom of the parameters (in particular, when the latter are close to zero). For instance, in the case of parameter L , the minimum of the new interval is $\min\left(\frac{L_{min}}{2}, L_{best} - \frac{L_{delta}}{2 \times n1}\right)$, where L_{min} and L_{max} are the minimum and maximum of the previous interval, $L_{delta} = L_{max} - L_{min}$, L_{best} is the L coordinate of the best parameter set, and $n1$ is the reduction parameter selected by the user. Thus, the minimum value of the test interval is divided by 2 at each time the best point of the grid is close enough to zero. The maximum value of the new interval is, more simply, $L_{best} - \frac{L_{delta}}{2 \times n1}$. These formulas, written for L , are also valid for β , μ , and p .

¹ Available at <https://github.com/SergeMOULIN/retrotransposons-spread>

3.2.2 Location in the genome

Formulas written for L also work for X , except that X cannot get out of the interval created by the lowest and highest positions of copies present in the observed genome. In addition, if X test interval no longer contains any copy position corresponding to those of the observed genome, the nearest from X_{best} is sought. Then X_{best} definitely takes the value of this position and is no longer estimated. The estimate continues on the other four parameters only.

3.3 Module and package dependencies

Hungarian method has been applied using the “munkres” module, implemented in 2008 by Brian M. Clapper (Clapper, 2008). Furthermore, kernel density estimations used in Part 4.3 to estimate T_{obs} have been fitted with R software: the function “density” of the package “stat”² has been used with all default settings, except the number of equally spaced points where the density is estimated ($n = 4096$ instead of 512).

4 Results and Discussion

4.1 The data

This proposal has been applied to the spread of the LTR retrotransposons ROO, DM412, and GYPSY on the chromosome 3L of the *Drosophila melanogaster* genome. This sequence corresponds to the left arm of the chromosome 3, which is the largest autosomal chromosome of *D. melanogaster*. This is also the most prolific chromosome for each of the LTR retrotransposons we considered, this is why it has been chosen for this case study. ROO corresponds to the LTR retrotransposon in *Drosophila melanogaster* with the largest number of copies (Kaminker *et al.*, 2002; Lerat *et al.*, 2003; De la Chaux and Wagner, 2009). DM412 is supposed to have been recently acquired by the *D. melanogaster* through horizontal transfer from a close relative species (Bartolomé *et al.*, 2009; Lerat *et al.*, 2011; Modolo *et al.*, 2014). Finally, GYPSY is an older and likely well regulated LTR retrotransposon (Lerat *et al.*, 2011). Chromosome 3L contains 32 copies of ROO (with a mean nucleotidic identity of 96%), 16 copies of DM412 (mean nucleotidic identity of 88.29%), and 6 of GYPSY (mean nucleotidic identity of 63.84%).

Two databases have been used during the experiments. The first one contains positions and nucleotidic sequences for each TE copy annotated in the *D. melanogaster* (flybase website³ version number 5 of the *D. melanogaster* genome (Adams *et al.*, 2000; Smith *et al.*, 2007)). The second database has been downloaded from the RepBase website⁴ and contains the consensus sequences for each TE corresponding to reference active elements. These reference sequences have been placed at the root of the TE tree, corresponding to the sequence state at time 0. Indeed, in the case of ROO, the chromosome 3L contains two copies 100% identical to the reference sequence, which cannot arrive by chance. This presence justifies to put such a sequence at the root of the tree (however, this is not the case for the two other TEs, that is why an ancestral reconstruction stage should be added to our algorithm, in order to set the root).

Then the Needleman-Wunsch distance between each TE copy (from the first database) and its reference (from the second database) has been calculated, in order to obtain the deterioration states.

In this case study, the estimation method described in Section 2.2.2 has actually been applied not only once but 60 times in each situation, in order to check the consistency of the obtained parameter sets. Then,

Table 3. Setting table

parameter	starting point	end point	interval division	desired accuracy
X_0	0	1	3	10^{-3}
μ	0.1	10	4	10^{-2}
β	0.01	1	3	10^{-3}
p	0.1	100	4	10^{-1}
L	0.01	1	3	10^{-3}

Table 4. Results

parameter	ROO	DM412	GYPSY
X_0	0.024	0.652	0.728
μ	3.072	0.447	0.119
β	0.016	0.0023	0.028
p	0.185	0.756	0.014
L	0.435	1.516	8.400×10^{-4}

in each situation, the best set has been determined by the minimization of $S = \sum_{i=1}^{20,000} D_T(T_i, G)$. The best parameter set of each case study is presented in Section 4.3. The whole obtained parameter sets are presented in supplementary data with their descriptive statistics. Some indications about consistency of these results are provided in Section 4.4.

4.2 Settings

Let us first recall that X_0 , which represents the root position in the chromosome, is inside the interval $[0, 1]$. In other words, copy positions have been divided by the chromosome size. For the chromosome 3L, this size has been set at 32,600,000 base pairs (bp) according to the sum of euchromatic and heterochromatic region lengths (Adams *et al.*, 2000).

In Table 3, each row represents the beginning and the end of the test interval, the number by which the test interval has been divided, and the final desired accuracy regarding the parameter. In particular, in the third column, the value associated to X_0 , β , and p is 3. This latter means that these parameters have been tested at the beginning, in the first third, in the second one, and at the end of the test interval. The value associated to μ and p is 4 (thus these parameters have been tested in 5 points). Finally, at each iteration, the grid contains $4^3 \times 5^2 = 1,600$ points. μ and p have been evaluated more than the other parameters, as it has been noted in our first analyses that these two parameters have a worse consistency than the other ones.

The other parameters are:

- $n1 = 1.5$: at each step, after the best point has been found, the grid’s dimensions have been divided by 1.5.
- $N1 = 100$: each point has been tested 100 times.
- $N2 = 100$: the 100 best points have been re-checked.
- $N3 = 1000$: the best points have been checked again 1000 times.

4.3 Results

The obtained parameters are summarized in Table 4.

If we consider, for instance, the ROO spread, the obtained parameters can be interpreted as follows:

- $X_0 = 0.0243473400$. Root position in chromosome 3L is $0.0243473400 \times 32,600,000 = 793,723$. The root is the copy located around the 793,723th nucleotide.

² Kernel Density Estimation in R, <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/density.html>

³ <http://flybase.org/>

⁴ <http://www.girinst.org/reprbase/>

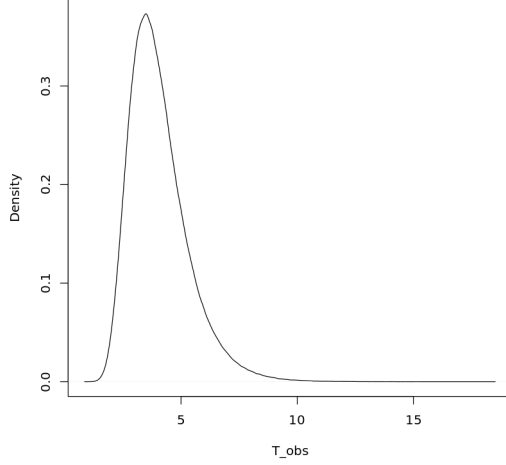


Fig. 3. Kernel estimation of T_{obs} for ROO

- $\mu = 3.072$. The average time between two mutations is 3.072, where 1 is the average time before duplication of the root. Mutations are thus less frequent than duplications. Please note that this estimation of μ is without time unit: it is related to the duplication speed of the root. It allows to estimate duplication speed when the mutation speed is known, and *vice versa*.
- $\beta = 0.016$. The proportion of nucleotides affected by a mutation follows a $\min(1, \mathcal{E}(\frac{1}{0.016}))$ distribution.
- $p = 0.185$. p allows to determine how many mutations led to a decrease in the duplication speed. For example, in this case, if the identity between a copy and the reference is 0.75 (*i.e.*, state of deterioration = 0.25), then the duplication speed of this copy is reduced by 4.625% (indeed $0.25 \times 0.185 = 0.04625$).
- $L = 0.435$. The distance traveled by the TE before insertion follows a distribution $E(\frac{1}{0.435})$, with the constraint that the new copy must remain on the chromosome (*cf.* Section 3.1.3).

The fact that some of the obtained parameters are outside the test interval chosen at the beginning of the program (for instance, $\beta = 2.32 \times 10^{-3}$ in DM412 case) is a desired effect, to let a larger freedom to the parameters. In particular, the aim was to let parameters to be as close as possible to zero if required (*cf.* Section 3).

In each of these three cases, one billion of trees have been simulated with the obtained parameter set. The best of these trees is shown in Figure 1 for ROO and in supplementary data for DM412 and GYPSY. Moreover, each simulation returns a propagation time, denoted by T_{obs} . Thus, this billion of simulations allowed us to draw histograms or kernel density estimations (Epanechnikov, 1969) of these propagation times. Here, Gaussian kernel density estimations (Cleveland and Devlin, 1988) have been represented on (*cf.* Figure 3 for ROO and in supplementary data for DM412 and GYPSY). Propagation times having the highest density are shown in Table 5. As for μ , this estimation of T_{obs} is without time unit, but it is related to the duplication speed of the root.

4.3.1 Focus on the roots

According to this model and method, the ROO root could correspond to the FBti0020009 copy. It is an incomplete copy (428 bp, compared to the reference which is 9112 bp length) and corresponds to a solo-LTR, a remnant from a LTR-LTR recombination. This copy is thus no longer

Table 5. Most likely values for T_{obs}

Transposable element	Most likely value for T_{obs}
ROO	3.490
DM412	2.705
GYPSY	1.671

Table 6. Consistency indicators

	X	μ	β	p	L
ROO	0.092	0.223	0.013	0.162	0.047
DM412	0.044	0.349	0.024	0.358	0.180
GYPSY	0.199	0.0047	0.012	5.290×10^{-4}	0.0016

active. The DM412 root could correspond to the FBti0020033 copy. This is a complete copy (7440 bp), which displays 98.8% identity with the reference. Its two LTRs are identical, indicating that its insertion into the genome is very recent. It is potentially still active since it possess two intact open reading frames encoding the genes gag and pol. Finally, GYPSY root could correspond to the FBti0060418 copy. It is an incomplete sequence (976 bp compared to the reference which is 7471 bp length) that is very divergent to the reference (78.45% identity). This copy corresponds to a piece of the inner part of the gypsy element and represents a very old and degraded copy that is not currently active.

4.4 Consistency of results

As explained previously, the optimization method has been actually applied 60 times for each TE. The descriptive statistics for these three cases are summarized in the supplementary data. In addition, quotients $\frac{\text{Standard deviation of the results}}{\text{Test interval}}$ have been calculated for each parameter, in each case, in order to assess the consistency of the results. These quotients are reproduced in Table 6.

4.5 Investigating other experimental setups

In this case study, five parameters (X , μ , β , p , and L) have been estimated in a first step. Then, T_{obs} has been deduced based on its kernel density estimation. Another approach have been implemented that estimate six parameters together at once. This approach can be summarized as follows:

- Instead of stopping the tree simulation according to the number of produced copies, the stop condition is $T_{obs} = 1$.
- The duplication speed is not fixed at 1 anymore but is a parameter denoted q to estimate. The six parameters to estimate are thus X , μ , β , q , p , and L .
- In this approach, the simulated trees will not necessarily have the same number of copies as the observed genome. A penalty for the difference of copy number can be added to the Kuhn-Munkres result.

The problem with this approach was that when q is a variable, it is really consistent while any other parameter does not really matter (they are like random). As an illustration, consistency indicators for this approach applied to ROO are presented in Table 7, Row 1. As can be seen in this table, the standard deviations of each of these parameters, as a proportion of the associated interval, is very large, while it should be as close as 0 in a consistent scenario.

Table 7. Consistency indicators for the first approaches in the ROO case

	X	μ	β	q	p	L
Approach 1	0.185	0.382	0.174	0.013	0.362	0.306
Approach 2	0.112	0.408	0.255	0.014	0.473	0.300
Final Approach	0.079	0.204	0.013	NA	0.344	0.166

We have also tried to estimate the whole $(X, \mu, \beta, q, p, \text{ and } L)$ tuple by minimising:

$$\sum_{i=1}^N D_T(T_i, G) \mathbf{1}_{\text{number of copies simulated} = \text{number of copies in observed genome}},$$

where $\mathbf{1}_A$ is the characteristic function of the subset A . Consistency indicators of this approach applied to ROO are shown in the second row of Table 7. For the sake of comparison, these consistency indicators are also given in the case of the approach finally adopted, in the last row of Table 7, and with the same computation time. Obviously, the latter provide the best results in terms of consistency. Note that, to achieve this comparison, we have enlarged the precision ($N1$, $N2$, and $N3$ have been increased), which explains why the results contained in the last row of Table 7 are not the same than what has been previously presented (Table 6, Row 1).

4.6 Conclusion and future perspectives

In this article, a model has been proposed for the propagation of LTR retrotransposons in a genome. Various functions have been implemented to simulate this spread as well as graphic representations. Finally, a first method for estimating the parameters of this propagation model has been proposed and applied to the spread of TEs on the ROO, GYPSY, and DM412 elements in a chromosome of *Drosophila melanogaster*.

This work can be improved in various directions, some of them being listed below.

Firstly, it was assumed that an unique root created every copies in the chromosome. The possibility of several roots can be considered too. For instance, a method of unsupervised classification like Gaussian Mixture model could be applied in order to detect the number of clusters.

Secondly, a deletion parameter, which models the disappearance of TE copies, could be considered too.

Furthermore all the parts of the genome are not equally favourable to the TE fixation due their deleterious effects on the neighboring sequences, which result in the elimination of certain insertion by negative selection. It may be appropriate to also take this fact into account.

The effect of an epigenetic regulation that can affect ET behaviour even if they do not face sequence degradation could be taken into account.

As precised in part 4.1, an ancestral reconstruction stage should be added to our algorithm, in order to set the sequence state of the root at time 0.

Finally, it would be useful to search approaches that allow a faster and more consistent estimation.

Computations have been performed on the supercomputer facilities of the Mésocentre de calcul de Franche-Comté.

Funding

This work has been supported by the Transposable Elements project of the Franche-Comté region.

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., et al. (2000). The genome sequence of drosophila melanogaster. *Science*, **287**(5461), 2185–2195.

Bartolomé, C., Bello, X., and Maside, X. (2009). Widespread evidence for horizontal transfer of transposable elements across drosophila genomes. *Genome Biol*, **10**(2), R22.

Belancio, V. P., Hedges, D. J., and Deininger, P. (2008). Mammalian non-ltr retrotransposons: for better or worse, in sickness and in health. *Genome research*, **18**(3), 343–358.

Clapper, B. (2008). munkres 1.0.7 for python. <https://pypi.python.org/pypi/munkres/>.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**(403), 596–610.

De la Chaux, N. and Wagner, A. (2009). Evolutionary dynamics of the ltr retrotransposons roo and rooa inferred from twelve complete drosophila genomes. *BMC evolutionary biology*, **9**(1), 205.

Drakos, N. E. and Wahl, L. M. (2015). Extinction probabilities and stationary distributions of mobile genetic elements in prokaryotes: The birth–death–diversification model. *Theoretical population biology*, **106**, 22–31.

Egner, C. and Berg, D. E. (1981). Excision of transposon tn5 is dependent on the inverted repeats but not on the transposase function of tn5. *Proceedings of the National Academy of Sciences*, **78**(1), 459–463.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, **14**(1), 153–158.

Foster, T. J., Lundblad, V., Hanley-Way, S., Halling, S. M., and Kleckner, N. (1981). Three tn10-associated excision events: relationship to transposition and role of direct and inverted repeats. *Cell*, **23**(1), 215–227.

Green, M. M. (1988). Mobile dna elements and spontaneous gene mutation. *Banbury Rep*, **30**, 41–50.

Kaminker, J. S., Bergman, C. M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D. A., Lewis, S. E., Rubin, G. M., et al. (2002). The transposable elements of the drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol*, **3**(12), RESEARCH0084.

Kaplan, N., Darden, T., and Langley, C. H. (1985). Evolution and extinction of transposable elements in mendelian populations. *Genetics*, **109**(2), 459–480.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, **2**(1-2), 83–97.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Lerat, E., Rizzon, C., and Biémont, C. (2003). Sequence divergence within transposable element families in the drosophila melanogaster genome. *Genome research*, **13**(8), 1889–1896.

Lerat, E., Burtet, N., Biémont, C., and Vieira, C. (2011). Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. *Gene*, **473**(2), 100–109.

McClintock, B. (1987). *The discovery of characterization of transposable elements: the collected papers of Barbara McClintock*.

Modolo, L., Picard, F., and Lerat, E. (2014). A new genome-wide method to track horizontally transferred sequences: application to drosophila. *Genome biology and evolution*, **6**(2), 416–432.

Moody, M. E. (1988). A branching process model for the evolution of transposable elements. *Journal of mathematical biology*, **26**(3), 347–357.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, **5**(1), 32–38.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, **48**(3), 443–453.

Sanmiguel, P. and Bennetzen, J. L. (1998). Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Annals of Botany*, **82**(suppl 1), 37–44.

Sawyer, S. A., Dykhuizen, D. E., DuBose, R. F., Green, L., Mutangadura-Mhlanga, T., Wolczyk, D. F., and Hartl, D. L. (1987). Distribution and abundance of insertion sequences among natural isolates of escherichia coli. *Genetics*, **115**(1), 51–63.

Smith, C. D., Shu, S., Mungall, C. J., and Karpen, G. H. (2007). The release 5.1 annotation of drosophila melanogaster heterochromatin. *Science*, **316**(5831), 1586–1591.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, **8**(12), 973–982.

Table 8. ROO, summary of the 60 results

	X	μ	β	p	L
Min.	0.0035	0.332	4.941×10^{-4}	0.185	0.339
1st Qu.	0.0035	1.100	0.0032	6.851	0.416
Median	0.030	2.417	0.0093	16.271	0.437
Mean	0.057	2.937	0.013	19.963	0.437
3rd Qu.	0.075	3.612	0.021	30.747	0.461
Max.	0.211	8.886	0.052	79.664	0.586

Table 9. DM412, summary of the 60 results

	X	μ	β	p	L
Min.	0.631	0.084	2.493×10^{-4}	0.7562	0.954
1st Qu.	0.652	0.327	0.0012	15.432	1.091
Median	0.652	0.505	0.0021	53.094	1.250
Mean	0.666	2.334	0.012	49.176	1.238
3rd Qu.	0.652	1.732	0.0088	71.831	1.373
Max.	0.723	11.969	0.104	151.412	1.614

Table 10. GYPSY, summary of the 60 results

	X	μ	β	p	L
Min.	0.715	0.074	0.017	0.0011	2.776×10^{-5}
1st Qu.	0.724	0.123	0.029	0.017	0.0012
Median	0.724	0.150	0.036	0.031	0.0022
Mean	0.725	0.155	0.037	0.049	0.0024
3rd Qu.	0.728	0.188	0.044	0.059	0.0031
Max.	0.730	0.321	0.075	0.305	0.0069

**Supplementary data related to the article
“Simulation based estimation of branching
models for retrotransposons”**

This documents provide supplementary data related to the article “Simulation based estimation of branching models for retrotransposons”. As explained in the main article, the estimation method has been used 60 times in each situation (*i.e.*, in ROO, DM412, and GYPSY case). The descriptive statistics for these 60 trials are summarized in Table 8 for ROO case, Table 9 for DM412 case, and Table 10 for GYPSY case. In addition, the trees that represent DM412 and GYPSY spreads are drawn in Figure 4 and Figure 6 respectively. Gaussian kernel density estimations of propagation times for DM412 and GYPSY are finally presented in Figure 5 and 7 respectively.

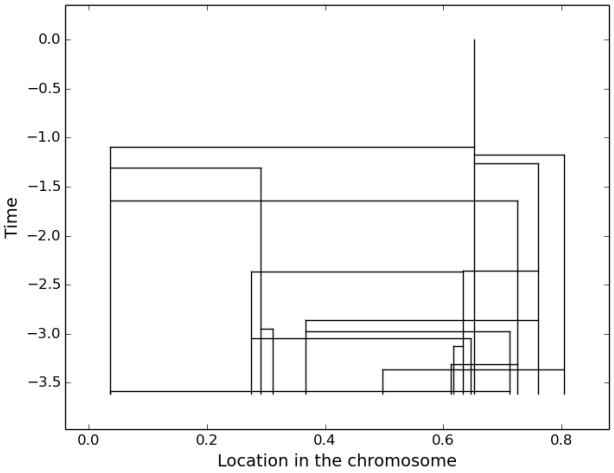


Fig. 4. DM412 spread

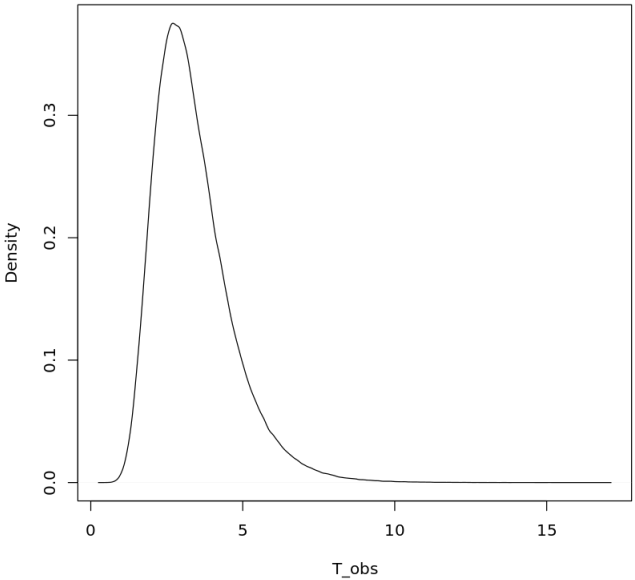


Fig. 5. Kernel estimation of T_{obs} for DM412

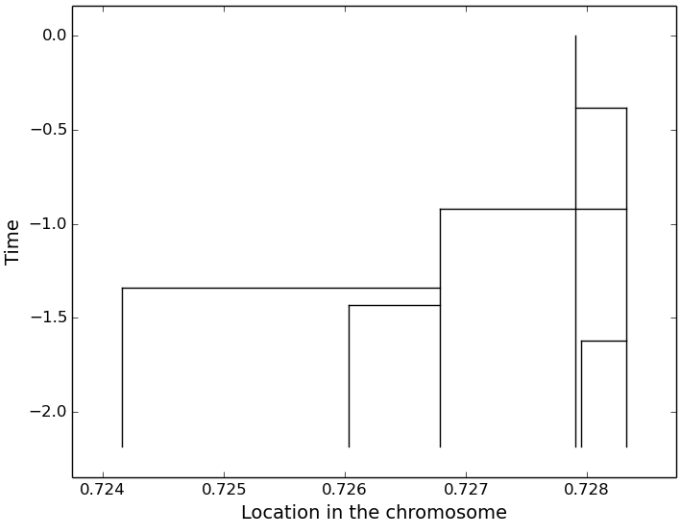


Fig. 6. GYPSY spread

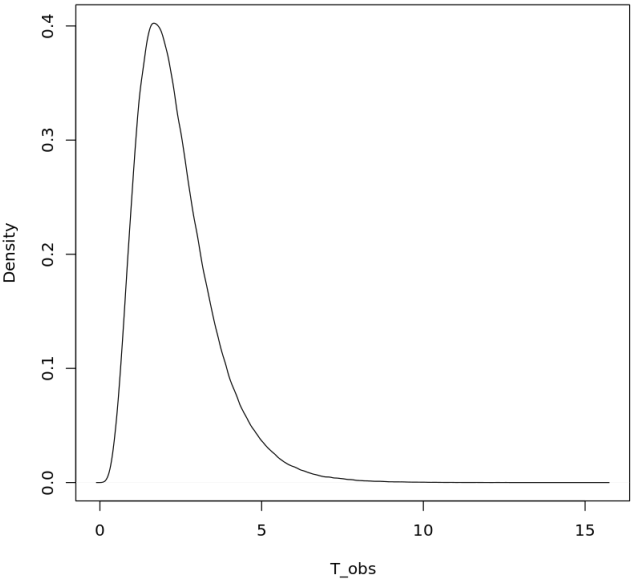


Fig. 7. Kernel estimation of T_{obs} for GYPSY